

# LARGE DEVIATIONS AND SUM RULES FOR SPECTRAL THEORY

## A PEDAGOGICAL APPROACH

JONATHAN BREUER<sup>1,4</sup>, BARRY SIMON<sup>2,5</sup>,  
AND OFER ZEITOUNI<sup>3,6</sup>

**ABSTRACT.** This is a pedagogical exposition of the large deviation approach to sum rules pioneered by Gamboa, Nagel and Rouault. We'll explain how to use their ideas to recover the Szegő and Killip–Simon Theorems. The primary audience is spectral theorists and people working on orthogonal polynomials who have limited familiarity with the theory of large deviations.

### 1. INTRODUCTION AND SUM RULES

This note discusses a new approach to sum rules in the spectral theory of orthogonal polynomials on the unit circle (OPUC) and real line (OPRL). We are given a probability measure,  $d\mu$ , on  $\partial\mathbb{D}$  or  $\mathbb{R}$  of the form:

$$d\mu(\theta) = w(\theta)\frac{d\theta}{2\pi} + d\mu_s(OPUC); \quad d\mu(x) = w(x)dx + d\mu_s(OPRL) \quad (1.1)$$

where  $d\mu_s$  is singular w.r.t  $d\theta$  or  $dx$ . The recursion relations for OPUC are given for the monic OPs,  $\{\Phi_n\}_{n=0}^\infty$ , by

$$\Phi_{n+1}(z) = z\Phi_n(z) - \overline{\alpha_n}\Phi_n^*(z); \quad \Phi_0 \equiv \mathbf{1}; \quad \Phi_n^*(z) = \overline{z^n\Phi_n\left(\frac{1}{\bar{z}}\right)} \quad (1.2)$$

---

*Date:* December 2, 2016.

*2010 Mathematics Subject Classification.* 60F10, 35P05, 42C05.

*Key words and phrases.* sum rules, large deviations, orthogonal polynomials.

<sup>1</sup> Institute of Mathematics, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel. E-mail: jbreuer@math.huji.ac.il.

<sup>2</sup> Departments of Mathematics and Physics, Mathematics 253-37, California Institute of Technology, Pasadena, CA 91125. E-mail: bsimon@caltech.edu.

<sup>3</sup> Faculty of Mathematics, Weizmann Institute of Science, POB 26, Rehovot 76100, Israel and Courant Institute, NYU. E-mail: ofer.zeitouni@weizmann.ac.il.

<sup>4</sup> Research supported in part by Israeli BSF Grant No. 2014337.

<sup>5</sup> Research supported in part by NSF grant DMS-1265592 and in part by Israeli BSF Grant No. 2014337.

<sup>6</sup> Research supported in part by a grant from the Israel Science Foundation.

with Verblunsky coefficients  $|\alpha_n| \leq 1$ , and for OPRL and the orthonormal polynomials,  $\{p_n\}_{n=0}^\infty$ , by

$$xp_n(x) = a_{n+1}p_{n+1}(x) + b_{n+1}p_n(x) + a_np_{n-1}(x); \quad p_{-1} \equiv 0 \quad (1.3)$$

with Jacobi parameters  $b_n \in \mathbb{R}$ ,  $a_n \geq 0$ .

Verblunsky's Theorem (see [36, 39]) says that there is a bijection  $V : \prod_{n=0}^\infty \mathbb{D} \rightarrow \{\mu \text{ on } \partial\mathbb{D} \mid \mu(\partial\mathbb{D}) = 1; \text{ support of } \mu \text{ is infinite}\}$  that maps the Verblunsky coefficients,  $\{\alpha_n\}_{n=0}^\infty$ , to a measure with those Verblunsky coefficients. Here  $\mathbb{D}$  will denote the open unit disc in the complex plane,  $\mathbb{C}$ , and we say the support is infinite if it is not a finite set of points. Similarly, for each  $n$ , there is a bijection,  $V_n$ , of  $\left(\prod_{j=0}^{n-2} \mathbb{D}\right) \times \partial\mathbb{D}$  to probability measures on  $\partial\mathbb{D}$  with exactly  $n$  points in their support (i.e.  $\alpha_j \in \mathbb{D}, j = 0, \dots, n-2; \alpha_{n-1} \in \partial\mathbb{D}$ .) Moreover, the maps are homeomorphisms if the  $\alpha$ 's are given the topology of point-wise convergence (product topology) and the measures the topology of vague convergence (i.e. weak topology as functionals on all continuous functions).

In some sense, it is natural to consider all measures at once and the corresponding set of possible Verblunsky coefficient sequences. On this larger set, the topology is no longer a product topology (which it can't be since it is a union of products but not a product itself). A Verblunsky coefficient sequence is a sequence  $\alpha = \{\alpha_n\}_{n=0}^{N(\alpha)-1}$  with  $N(\alpha) \leq \infty$ . If  $N(\alpha) < \infty$  the sequence has  $N(\alpha)$  elements with the last one in  $\partial\mathbb{D}$  and the others in  $\mathbb{D}$ . The topology is metrizable and  $\alpha^{(j)}$  converges to  $\alpha^{(\infty)}$  by the following rule: If  $N(\alpha^{(\infty)}) = \infty$ , then we require that  $N(\alpha^{(j)}) \rightarrow \infty$  and  $\alpha_k^{(j)} \rightarrow \alpha_k^{(\infty)}$  for each fixed  $k$  as  $j \rightarrow \infty$ . If  $N(\alpha^{(\infty)}) = N_0 < \infty$ , we require that  $\alpha_k^{(j)} \rightarrow \alpha_k^{(\infty)}$  for  $k = 0, \dots, N_0 - 1$ . In this topology the space of Verblunsky coefficients sequences is compact and  $V$  maps this space homeomorphically onto the space of all probability measures with the weak topology. This shows also that the topology on Verblunsky coefficients is metrizable.

Similarly, Favard's Theorem (see [39]) says that there is a bijection,  $J : \prod_{n=1}^\infty (\mathbb{R}, (0, \infty))$  with  $\sup(|a_n| + |b_n|) < \infty \rightarrow \{\mu \text{ on } \mathbb{R} \mid \mu(\mathbb{R}) = 1, \text{ supp } \mu \text{ is infinite and bounded}\}$  that maps the Jacobi parameters  $\{a_n, b_n\}_{n=1}^\infty$  to the measure whose recursion parameters in (1.3) are the given Jacobi parameters. Again, there are bijections,  $J_n$ , mapping  $\{a_j\}_{j=1}^{n-1} \cup \{b_j\}_{j=1}^n$  to  $n$ -point measures. The question of continuity is a little subtle. We note here that restricted to measures with support in  $[-R, R]$  and the corresponding set of  $a$ 's and  $b$ 's, the map is a homeomorphism when the measures are given the weak topology and the Jacobi parameters the topology described in the last paragraph (with

$\alpha_{N-1} \in \mathbb{D}$  replaced by  $a_N = 0$ ). We'll say more about the issue of topologies in (g) of Section 5.

We'll focus here on two of the simplest and most basic sum rules. One is the Szegő–Verblunsky sum rule for OPUC, basically Szegő's Theorem in the form written down by Verblunsky [43] (who was the first to prove the Theorem when  $d\mu_s \neq 0$ ):

$$\sum_{j=0}^{\infty} -\log(1 - |\alpha_j|^2) = - \int \log(w(\theta)) \frac{d\theta}{2\pi} \quad (1.4)$$

Critical for applications to spectral theory, all terms in the sum are positive so one can prove and interpret this in all cases even when both sides are infinite. In particular, the fact that two sides are finite simultaneously implies that

$$\sum_{j=0}^{\infty} |\alpha_j|^2 < \infty \iff \int \log(w(\theta)) \frac{d\theta}{2\pi} > -\infty \quad (1.5)$$

[39] uses the term ‘spectral theory gem’ for a result like this where there is a strict equivalence between conditions on the spectral data (i.e. the measure) and conditions on the recursion coefficients.

The other is the sum rule of Killip–Simon [26] for OPRL:

$$Q(\mu) + \sum_{\mu(\{E_n\}) > 0, |E_n| > 2} F(E_n) = \sum_{n=1}^{\infty} \left[ \frac{1}{4} b_n^2 + \frac{1}{2} G(a_n) \right] \quad (1.6)$$

where

$$Q(\mu) = \frac{1}{4\pi} \int_{-2}^2 \log \left( \frac{\sqrt{4-x^2}}{2\pi w(x)} \right) \sqrt{4-x^2} dx \quad (1.7)$$

$$F(E) = \frac{1}{2} \int_2^{|E|} \sqrt{x^2 - 4} dx \quad (1.8)$$

$$G(a) = a^2 - 1 - \log(a^2) \quad (1.9)$$

with the condition that  $Q(\mu) = \infty$  unless  $\{x \mid w(x) \neq 0\} = [-2, 2]$  (up to sets of Lebesgue measure zero).

Again, all terms in (1.6) are positive. Moreover  $G(a) = 2(a-1)^2 + O((a-1)^3)$ ,  $F(E) = \frac{2}{3}(|E|-2)^{3/2} + O((|E|-2)^{5/2})$  which means that

(1.6) implies the gem (Killip–Simon Theorem) that

$$\sum_{n=1}^{\infty} (a_n - 1)^2 + b_n^2 < \infty$$

$$\iff$$

$$\text{ess supp } (d\mu) = [-2, 2], \quad Q(\mu) < \infty \text{ and } \sum_m (|E_m| - 2)^{3/2} < \infty$$
(1.10)

where  $\text{ess supp } (d\mu)$  is the support of  $d\mu$  with isolated atoms removed.

There are many proofs of Szegő's Theorem and so of (1.4) (see Chapter 2 of [36]) but until recently, essentially one proof of the Killip–Simon sum rule – the original proof of [26] or variants (e.g. [35] or [39]). There is a simple analog of this proof for Szegő's Theorem (see [39, Section 2.7]).

This proof starts by showing a so-called step-by-step sum rule. One then uses positivity of the terms and semi-continuity of the entropy (an idea appearing already in Verblunsky [43]) to get the full sum rule. The step-by-step rules come from a Jensen formula (e.g. [41, Section 9.8]) for suitable functions on the disk. The value and derivatives at zero involve recursion coefficients and the integral over  $\partial\mathbb{D}$  the measure side. By taking derivatives at 0 of a Poisson–Jensen formula, one actually gets an infinite set of step-by-step sum rules related to work of Case [8] in the Jacobi matrix situation. The sum rule involving the  $n$ th derivative is called  $C_n$  by Killip–Simon. None of the  $C_n$  sum rules have positive terms but Killip–Simon discovered that  $C_0 + \frac{1}{2}C_2$  has positive terms and that yielded (1.6). For a long time, there was no explanation for why this turned out to be positive other than good luck. Work of Nazarov et al [33] (see Denissov–Kupin [13] for OPUC) shed some light on the positivity condition but it was still unclear why certain combinations of the Taylor coefficients involved are positive. Moreover, the functions involved in the sum rules (e.g.,  $Q$ ,  $F$ , and  $G$  of (1.7)–(1.9) above) seem to simply ‘pop out’ of these combinations and their significance was generally unclear.

Recently Gamboa, Nagel and Rouault [17] (henceforth GNR; they rely on an earlier work of two of those authors [20] and have further papers [18, 19]) changed this situation. Using the theory of large deviations, they found a new proof of the Szegő–Verblunsky and Killip–Simon sum rules. One punch line of their work is that the Szegő–Verblunsky sum rule follows from a large deviation principle for the spectral measures of the Circular Unitary Ensemble (CUE) of random matrix theory and that the Killip–Simon sum rule follows from

a large deviation principle for the spectral measures of the Gaussian Orthogonal Ensemble (GOE). The paper [17] is written for experts in probability theory and is somewhat opaque to spectral theorists. This expository paper is intended to give spectral theorists and people working on orthogonal polynomials access to this work of importance to them. It may also be useful to probabilists who want to understand the context of the results.

Ignoring technical issues for now, a family of measures,  $\{d\mathbb{P}_N\}_N$  on a space,  $X$ , is said to obey a large deviation principle (LDP) if the probability that  $x$  is near  $x_0$  goes roughly as  $e^{-NI(x_0)}$  as  $N \rightarrow \infty$ .  $I(x)$  is called the rate function and is necessarily non-negative. Suppose now that  $\{d\mathbb{P}_N\}$  is a family of probability measures on the set of probability measures on  $\partial\mathbb{D}$ . Then the bijection  $V$  drags these measures to a set of measures,  $\widetilde{d\mathbb{P}_N}$ , on the set of possible Verblunsky coefficients (i.e. on  $\prod_{j=0}^{\infty} \mathbb{D} \cup \bigcup_{k=1}^{\infty} (\mathbb{D}^{k-1} \cup \partial\mathbb{D})$ ). Since LDP is defined in terms of “nearby” and  $V$  is a homeomorphism, the  $d\mathbb{P}_N$  obey a LDP with rate function,  $I_M(\mu)$ , if and only if the  $\widetilde{d\mathbb{P}_N}$  obey a LDP with rate function  $I_V(\alpha)$  where

$$I_V(\alpha) = I_M(\mu) \quad \text{whenever } \mu = V(\alpha) \quad (1.11)$$

$V(\alpha)$  is the measure associated to  $\alpha$  so (1.11) is precisely an equality of a function of the recursion coefficients and a function of its associated measure, i.e. a sum rule. Moreover, since a rate function is necessarily non-negative, both sides of (1.11) are positive, so this is a way of generating *positive* sum rules.

This approach illuminates why the various quantities in the Killip–Simon sum rule, that in their proof occur by ad hoc considerations, arise naturally. The  $\sqrt{4-x^2}/2\pi$  in their quasi–Szegő condition (i.e., in  $Q$  of (1.7)) is just the Wigner semi–circle density for GOE. The function  $G$  is just the large deviations rate function for averages of exponential random variables (see (2.9) below) and the function  $F$  is just the external potential for Coulomb interaction in an external quadratic field (see (d) of Section 5).

In Section 2, we present a quick overview of the theory of large deviations. In Section 3, we’ll compute the measure side of the LDP that yields the Szegő–Verblunsky sum rule and in Section 4, the coefficient side. Section 5 describes both sides of the LDP that leads to the Killip–Simon sum rule and Section 6 has some remarks on further developments. The Appendix presents an inductive proof to a formula by Killip and Nenciu ([25] and Theorem 4.2 below) regarding the probability distribution of the Verblunsky coefficients of a Haar distributed unitary matrix.

We thank Peter Yuditskii for telling two of us (JB and BS) about [17] and Fritz Gesztesy for encouraging us to write a pedagogic note.

## 2. LARGE DEVIATIONS

Here we'll sketch some of the key ideas in the theory of large deviations. Two books on the subject are Deuschel–Stroock [14] and Dembo–Zeitouni [12]. As both books point out, there is not so much a theory as a collection of powerful tools. While the subject has roots going back to Laplace, the modern framework goes back to Varadhan, Donsker–Varadhan and Freidlin–Wentzel in the 1960's and 1970's.

Let  $\{\mathbb{P}_N\}_{N=1}^\infty$  be a sequence of probability measures on a Polish space,  $X$  (see Simon [40, Section 4.14] or Billingsley [6] for measure theory on Polish spaces). Let  $I$  be a non-negative function on  $X$  and  $\{a_N\}_{N=1}^\infty$  a sequence of positive numbers with  $a_N \rightarrow \infty$ . We say that  $\{\mathbb{P}_N\}_{N=1}^\infty$  obeys a LDP with rate function,  $I$ , and speed  $\{a_N\}_{N=1}^\infty$  if and only if

- (1)  $I$  is non-negative and lower semicontinuous on  $X$
- (2) For every open set,  $U \subset X$ , we have that

$$\liminf_{N \rightarrow \infty} \frac{1}{a_N} \log \mathbb{P}_N(U) \geq - \inf_{x \in U} I(x) \quad (2.1)$$

(3) For every closed set,  $K \subset X$ , we have that

$$\limsup_{N \rightarrow \infty} \frac{1}{a_N} \log \mathbb{P}_N(K) \leq - \inf_{x \in K} I(x) \quad (2.2)$$

We note that the rate function is uniquely determined by these conditions because lower semicontinuity implies that  $I(x_0) = \lim_B \inf_{x \in B} I(x)$  where the limit is over the directed set of all open (or over all closed) neighborhoods of  $x_0$ , directed by inverse inclusion. A rate function is called *good* if for each positive  $\alpha$ ,  $\{x \mid I(x) \leq \alpha\}$  is compact (note that by the lower semicontinuity, this set is always closed). The following two elementary results, whose proofs we leave to the readers, will be useful.

**Theorem 2.1.** *Let  $a_N = N^\ell$  (for  $\ell > 0$ ). Fix  $N_0$ . Then  $\{\mathbb{P}_{N+N_0}\}_{N=1}^\infty$  obeys a LDP with speed  $a_N$  and rate function,  $I$ , if and only if  $\{\mathbb{P}_N\}_{N=1}^\infty$  does.*

**Remark.** All that is needed is that  $\lim_{N \rightarrow \infty} \frac{a_{N+N_0}}{a_N} = 1$ .

**Theorem 2.2.** *Let  $U \subset \mathbb{R}^\nu$  be open. Let  $G$  be continuous on  $U$  with  $\lim_{x \rightarrow \partial U \cup \{\infty\}} G(x) = \infty$  and  $\inf_{x \in U} G(x) = 0$ . Let  $F \in L^1(\mathbb{R}^\nu)$  with  $\text{supp} F \subset \overline{U}$ ,  $F \geq 0$  and  $\inf_{x \in K} F(x) > 0$  for all compact  $K \subset U$ . Let  $d\mathbb{P}_N(x) = Z_N^{-1} e^{-NG(x)} F(x) d^\nu x$  where  $Z_N = \int e^{-NG(x)} F(x) d^\nu x$ . Then  $\mathbb{P}_N$  obeys a LDP with speed  $N$  and good rate function  $G$ .*

**Remark.** The assumptions are overly strong but suffice for the applications we make below.

A basic result is:

**Theorem 2.3** (Cramér's Theorem). *Given a real random variable,  $\xi$ , let*

$$\Lambda(\lambda) = \log \mathbb{E}(e^{\lambda\xi}) \quad (2.3)$$

*be its cumulant generating function and*

$$I(x) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \Lambda(\lambda)) \quad (2.4)$$

*its Legendre transform. Let  $\mathbb{P}_N$  be the probability distribution for  $N^{-1}S_N \equiv N^{-1}(X_1 + \cdots + X_N)$ , where  $\{X_j\}_{j=1}^\infty$  are independent copies of  $\xi$ . Then  $\mathbb{P}_N$  obeys a LDP with speed  $N$  and good rate function  $I$ .*

**Remark.** By Jensen's inequality,  $\Lambda$  is convex and, as a sup of linear functions, so is  $I(x)$ . If  $\Lambda$  is everywhere finite, it is  $C^1$ ,  $I(\bar{x}) = 0$  where  $\bar{x} = \mathbb{E}(\xi)$  and  $I(x) > 0$  for  $x \neq \bar{x}$ . Thus this result amplifies the law of large numbers.

*Sketch.* We sketch the proof in case  $\Lambda(\lambda) < \infty$  for all  $\lambda$  and the image of  $\xi$  is the whole real line; see [12] for the full result. Under these conditions, a bit of calculus and convex analysis shows that  $I(x)$  is non-negative, strictly convex, with  $I(\bar{x}) = 0$  where  $\bar{x} = \mathbb{E}(\xi)$ , and further that

$$I(x) = x\lambda_x - \Lambda(\lambda_x) \text{ with } \Lambda'(\lambda_x) = x \text{ and } \lambda_x \geq 0 \text{ iff } x \geq \bar{x} \quad (2.5)$$

(The key point in (2.5) is that a solution to the equation  $\Lambda'(\lambda_x) = x$  exists; it is mostly here that we use the assumptions that the image of  $\xi$  is  $\mathbb{R}$  and that  $\text{Dom}(\Lambda) = \mathbb{R}$ .)

To see the large deviations upper bound, we note first that the strict convexity of  $I(\cdot)$  implies that the latter is strictly monotone increasing (resp. decreasing) on  $[\bar{x}, \infty)$  (resp.  $(-\infty, \bar{x}]$ ). It is therefore enough to prove the upper bound on intervals of the form  $(-\infty, x]$  (with  $x < \bar{x}$ ) or  $[x, \infty)$  (with  $x > \bar{x}$ ). Considering the latter, we have, with  $\lambda \geq 0$ ,

$$\mathbb{P}(S_n/n \geq x) \leq \mathbb{E}(e^{-n\lambda x + \lambda S_n}) = e^{-n(\lambda x - \Lambda(\lambda))}$$

where the independence of the  $X_i$ 's was used in the last equality. Choosing  $\lambda = \lambda_x$  completes the proof of the large deviations upper bound.

To see the lower bound, it is enough to show that

$$\lim_{\delta \rightarrow 0} \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n/n \in (x - \delta, x + \delta)) = -I(x). \quad (2.6)$$

To see the latter, introduce the probability distribution  $\nu$  by setting  $d\nu/d\mathbb{P}_1(y) = e^{\lambda_x y - \Lambda(\lambda_x)}$ . Then,

$$\begin{aligned}\mathbb{E}_\nu(X_1) &= \int y e^{\lambda_x y - \Lambda(\lambda_x)} d\mathbb{P}_1(y) \\ &= e^{-\Lambda(\lambda_x)} \frac{d}{d\lambda} e^{\Lambda(\lambda)} \Big|_{\lambda=\lambda_x} \\ &= \Lambda'(\lambda_x) = x\end{aligned}$$

so, by (2.5),

$$\begin{aligned}\mathbb{P}(S_n/n \in (x - \delta, x + \delta)) &= \int_{\sum_{i=1}^n x_i/n \in (x - \delta, x + \delta)} \prod_{i=1}^n d\mathbb{P}_1(x_i) \\ &= \int_{\sum_{i=1}^n x_i/n \in (x - \delta, x + \delta)} e^{-\lambda_x \sum_{i=1}^n x_i + n\Lambda(\lambda_x)} \prod_{i=1}^n d\nu(x_i) \\ &\geq e^{-n(\lambda_x \delta + x\lambda_x - \Lambda(\lambda_x))} \int_{\sum_{i=1}^n x_i/n \in (x - \delta, x + \delta)} \prod_{i=1}^n d\nu(x_i) \\ &= e^{-n(\lambda_x \delta + I(x))} \int_{\sum_{i=1}^n x_i/n \in (x - \delta, x + \delta)} \prod_{i=1}^n d\nu(x_i)\end{aligned}\tag{2.7}$$

Since  $E_\nu(X_1) = x$ , the law of large numbers implies that for any  $\delta > 0$ , the last integral in (2.7) converges to 1 as  $n \rightarrow \infty$ . Taking the limits  $n \rightarrow \infty$  first and then  $\delta \rightarrow 0$  completes the proof of the lower bound.  $\square$

**Example 2.4.** Relevant to our considerations later is the average of exponential random variables. So let  $\{X_j\}_{j=1}^\infty$  be independent, identically distributed random variables (iidrv) with density  $\chi_{[0,\infty)}(x)e^{-x}dx$ . The cumulant generating function is

$$\Lambda(\lambda) = \log \left( \int_0^\infty e^{\lambda x} e^{-x} dx \right) = \begin{cases} -\log(1 - \lambda), & \text{if } \lambda < 1 \\ \infty, & \text{if } \lambda \geq 1 \end{cases} \tag{2.8}$$

For  $x \leq 0$ , taking  $\lambda \rightarrow -\infty$  in  $\lambda x - \Lambda(\lambda)$ , we see that  $I(x) = \infty$ . If  $x > 0$ , the  $\lambda$  derivative of  $\lambda x - \Lambda(\lambda)$  vanishes at  $\lambda = 1 - x^{-1}$  at which point  $\lambda x - \Lambda(\lambda)$  has the value  $x - 1 - \log(x)$ . Thus

$$\varphi(x) \equiv I(x) = \begin{cases} x - 1 - \log(x), & \text{if } x > 0 \\ \infty, & \text{if } x \leq 0 \end{cases} \tag{2.9}$$

is the (good) rate function. It is no coincidence as we'll see that the function of (1.9) is  $G(a) = \varphi(a^2)$ . We summarize the combination of



this calculation and Cramér's Theorem in the theorem below. The gamma distribution with parameters  $\alpha, \beta$  ( $\alpha, \beta > 0$ ) is the measure given by

$$dG_{\alpha, \beta}(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)} \chi_{[0, \infty)}(x) dx \quad (2.10)$$

For exponential iidrv,  $n^{-1} \sum_{j=1}^n X_j$  has distribution  $G_{n, n}$ , so this example allows one to also read off a LDP for suitable gamma distributions.

**Theorem 2.5.** *Let  $\ell_N$  be integers with*

$$\lim_{N \rightarrow \infty} N^{-1} \ell_N = \alpha > 0 \quad (2.11)$$

*Then  $Y_N \equiv N^{-1} \sum_{j=1}^{\ell_N} X_j$  with  $X_j$  iid exponential random variables obeys a LDP with speed  $N$  and rate function*

$$\varphi_\alpha(y) \equiv \alpha \varphi(y/\alpha) = y - \alpha - \alpha \log(y/\alpha) \quad (2.12)$$

**Remark.** This goes beyond the direct use of Cramér in two ways. First, we note that if real valued  $Z_N$  have a LDP with speed  $N$  and rate  $I$ , then  $\alpha Z_N$  has a LDP with speed  $\alpha N$  and rate  $\alpha I(\cdot/\alpha)$  by a trivial calculation. Secondly, if  $\alpha_N = \ell_N/N$  and  $\alpha_N \rightarrow 1$ , then  $\alpha_N^{-1} Y_N$  has a LDP with speed  $N$  and rate  $I$  if  $Y_N$  does and the rate function is continuous.

Next, we discuss a result known as the contraction principle which allows one to pull a LDP over under continuous maps. For our basic situation, the maps are homeomorphisms so it is trivial that LDP's carry over but in two places below we'll need the following:

**Theorem 2.6** (Contraction Principle). *Let  $X$  and  $Y$  be Polish spaces and  $f : X \rightarrow Y$  a continuous function onto  $Y$ . Suppose  $\{\mathbb{P}_N\}_{N=1}^\infty$  is a family of probability measures on  $X$  that obeys a LDP with speed  $a_N$  and good rate function  $I$ . Define on  $Y$  the function*

$$I^{(f)}(y) = \inf\{I(x) \mid f(x) = y\} \quad (2.13)$$

*Then the family of measures on  $Y$  defined by*

$$\mathbb{P}_N^{(f)}(A) = \mathbb{P}_N(f^{-1}[A]) \quad (2.14)$$

*obeys a LDP with speed  $a_N$  and good rate function  $I^{(f)}$ .*

*Proof.* A simple argument shows that  $I^{(f)}$  is a good rate function (see [12, Section 4.2.1]). If  $A$  is open (resp. closed), so is  $f^{-1}[A]$  and

$$\inf_{y \in A} I^{(f)}(y) = \inf_{x \in f^{-1}[A]} I(x) \quad (2.15)$$

so the LDP bounds for  $\mathbb{P}_N$  carry over to such bounds for  $\mathbb{P}_N^{(f)}$ .  $\square$

The last topic that we want to consider in this section is the theory of projective limits of LDP's or at least a very special case – projective limits are indexed by directed sets; we'll only consider the case where the directed set is  $\mathbb{Z}_+$ . We have Polish spaces  $\{X_j\}_{j=1}^\infty$  and  $X$  and continuous maps  $\pi_j : X \rightarrow X_j$  and  $\pi_{j+1,j} : X_{j+1} \rightarrow X_j$  all onto so that  $\pi_{j+1,j}\pi_{j+1} = \pi_j$ . We require that if  $\pi_j(x) = \pi_j(y)$  for  $x, y \in X$  and all  $j$ , then  $x = y$ . (In the abstract discussion, one just needs to be given  $X_j$  and  $\pi_{j+1,j}$  and can form  $X$  as the subset of  $\prod_{j=1}^\infty X_j$  of those  $x = (x_j)$  with  $\pi_{j+1,j}(x_{j+1}) = x_j$ . One puts the product topology on  $X$ . In the cases of interest,  $X$  will be explicitly given but it agrees with this abstract construction.) For a measure  $\mathbb{P}$  on  $X$ , define  $\pi_j^*(\mathbb{P})$ , a measure on  $X_j$  by  $\pi_j^*(\mathbb{P})(A) = \mathbb{P}(\pi_j^{-1}[A])$ . Here is a basic theorem due to Dawson–Gärtner [10] (see [12] for a proof):

**Theorem 2.7** (Projective Limit Theorem). *Let  $\{\mathbb{P}_N\}_{N=1}^\infty$  be a family of measures on  $X$ . Suppose that for each  $j$ ,  $\{\pi_j^*(\mathbb{P}_N)\}_{N=1}^\infty$  obeys a LDP with speed  $a_N$  and good rate function,  $I_j$  on  $X_j$ . Let*

$$I(x) = \sup_j \{I_j(\pi_j(x))\} \quad (2.16)$$

*Then  $I$  is a good rate function and  $\{\mathbb{P}_N\}_{N=1}^\infty$  obeys a LDP with speed  $a_N$  and rate function  $I$ .*

**Remarks.** 1. The converse, i.e. if  $\{\mathbb{P}_N\}_{N=1}^\infty$  obeys a LDP then so does each  $\{\pi_j^*(\mathbb{P}_N)\}_{N=1}^\infty$ , is trivial by the contraction principle.

2. The same idea shows that if  $\{\pi_{j+1}^*(\mathbb{P}_N)\}_{N=1}^\infty$  obeys a LDP, so does  $\{\pi_j^*(\mathbb{P}_N)\}_{N=1}^\infty$  and

$$I_j(x) = \inf \{I_{j+1}(y) \mid \pi_j(y) = x\} \quad (2.17)$$

which shows that  $I_j(\pi_j(x))$  is monotone in  $j$  so the sup in (2.16) is a limit.

**Example 2.8** ( $\mathbb{R}^\infty$ ). Take  $X_j = \mathbb{R}^j$ ,  $X = \mathbb{R}^\infty = \{(x_1, x_2, \dots) \mid x_j \in \mathbb{R}\}$  which is a Polish space and  $\pi_j(x)_k = x_k$  for  $k = 1, \dots, j$ . Theorem 2.7 says that to prove a LDP for  $X$ , we need only prove it for the finite dimensional  $\mathbb{R}^j$ .

**Example 2.9** ( $\mathcal{M}_{+,1}(\partial\mathbb{D})$ ). Let  $\mathbb{P}$  be a measure on  $\mathcal{M}_{+,1}(\partial\mathbb{D})$ , the probability measures on the unit circle. Given  $\mu \in \mathcal{M}_{+,1}(\partial\mathbb{D})$  and  $j = 1, 2, \dots$ , let  $\pi_j(\mu)$  be the point in  $\mathbb{R}^{2^j}$  with coordinates  $\mu(I_k^{(j)})$ ,  $k = 1, \dots, 2^j$  where

$$I_k^{(j)} = \{e^{2\pi i\theta} \mid \frac{k-1}{2^j} \leq \theta < \frac{k}{2^j}\} \quad (2.18)$$

Realizing  $\mathbb{R}^{2^j}$  as a set of measures, we can think of

$$\pi_j(\mu) = \sum_{k=1}^{2^j} \mu(I_k^{(j)}) 2^j \chi_{I_k^{(j)}}(x) dx \quad (2.19)$$

Thus  $\mathbb{P}$  induces a measure  $\pi_j^*(\mathbb{P})$  on either  $\mathbb{R}^{2^j}$  or on  $\mathcal{M}_{+,1}(\partial\mathbb{D})$  supported on a  $2^j$ -dimensional subspace. The  $\pi_j(\mu)$  determine  $\mu(\{e^{2\pi i\theta} \mid 0 \leq \theta < \frac{k}{2^j}\})$  and so  $\mu$ . Clearly as  $j \rightarrow \infty$ ,  $\pi_j(\mu)$  converges weakly to  $\mu$ .

In this case

$$\pi_{j+1,j}(y)_\ell = y_{2\ell-1} + y_\ell \quad \ell = 1, \dots, 2^j \quad (2.20)$$

Thus, to get a LDP for  $\mathcal{M}_{+,1}(\partial\mathbb{D})$ , we need only prove  $2^j$ -dimensional LDPs.

### 3. SZEGŐ'S THEOREM: MEASURE SIDE

We begin our presentation of the proof of Szegő's theorem using large deviations for CUE. CUE( $n$ ) is just another name for Haar measure on the unitary  $n \times n$  matrices. Let  $\{e_j\}_{j=1}^n$  be the standard basis for  $\mathbb{C}^n$ . It is easy to see that for a.e.  $U$ ,  $e_1$  is a cyclic vector for  $U$  so that  $U$  and  $e_1$  define a spectral measure

$$d\mu(\theta) = \sum_{j=1}^n w_j \delta_{\lambda_j} \quad (3.1)$$

on  $\partial\mathbb{D}$ , with precisely  $n$  pure points (aka atoms)  $\lambda_j = e^{i\theta_j}$ ,  $j = 1, \dots, n$ . Letting  $\{\varphi_j\}_{j=1}^n$  be the orthonormal basis of eigenvectors of  $U$ , so that  $U\varphi_j = \lambda_j\varphi_j$ , we have  $w_j = |\langle\varphi_j, e_1\rangle|^2$ . Of course, since  $\|e_1\| = 1$ ,

$$\sum_{j=1}^n w_j = 1 \quad (3.2)$$

Thus Haar measure induces a measure  $\mathbb{P}_n$  on  $n$ -point probability measures (which we can think of as a measure on all measures that happens to be supported on the  $n$ -point measures) which we also call CUE( $n$ ).

For  $\tilde{U}$  an arbitrary unitary,  $\tilde{U}U\tilde{U}^{-1}$  has the same eigenvalues as  $U$  and  $\langle\varphi_j(\tilde{U}U\tilde{U}^{-1}), e_1\rangle = \langle\tilde{U}\varphi_j(U), e_1\rangle$ . Since  $U \mapsto \tilde{U}U\tilde{U}^{-1}$  leaves Haar measure invariant, we see that the distribution of the unit vector  $(\langle\varphi_1(U), e_1\rangle, \langle\varphi_2(U), e_1\rangle, \dots, \langle\varphi_n(U), e_1\rangle) \in \mathbb{C}^n$  is invariant under unitary transformations, which implies it is the Euclidean measure restricted to the sphere. By using the fact that  $d^2z = \frac{1}{2}d\theta d(|z|^2)$  (which shows it is essential we work in  $\mathbb{C}$ ), it is not hard to show that

the squares of the components of a complex  $n$ -vector uniformly distributed on the sphere are uniformly distributed on the simplex (3.2). Thus we get that the  $\{w_j\}_{j=1}^n$  are independent of the eigenvalues and have  $\mathbb{P}_n$ -distribution

$$(n-1)! \chi_{\{\sum_{j=1}^{n-1} w_j \leq 1; w_j \geq 0\}}(w) dw_1 \dots dw_{n-1} \quad (3.3)$$

The distribution of the eigenvalues is given by the Weyl integration formula which says that the distribution of the eigenvalues under Haar measure is

$$\frac{1}{n!} |\Delta(e^{i\theta_1}, \dots, e^{i\theta_n})|^2 \prod_{j=1}^n \frac{d\theta_j}{2\pi} \quad (3.4)$$

$$\Delta(\lambda_1, \dots, \lambda_n) \equiv \prod_{i < j} (\lambda_i - \lambda_j) \quad (3.5)$$

For proofs of this formula from two different points of view, see Anderson et al [3, Section 4.17] or Simon [34, Section IX.3]. Thus

$$d\mathbb{P}_n(\theta_1, \dots, \theta_n, w_1, \dots, w_n) = \frac{1}{n(2\pi)^n} \chi_{\{\sum_{j=1}^{n-1} w_j \leq 1; w_j \geq 0\}}(w) |\Delta(e^{i\theta_1}, \dots, e^{i\theta_n})|^2 d\theta_1 \dots d\theta_n dw_1 \dots dw_{n-1} \quad (3.6)$$

With this in hand, we turn to the proof of the measure theory LD result:

**Theorem 3.1.**  *$\mathbb{P}_n$  as a family of measures on the set of probability measure on  $\partial\mathbb{D}$  obeys a LDP with speed  $n$  and good rate function*

$$I(\mu) = - \int \log w(\theta) \frac{d\theta}{2\pi} \quad (3.7)$$

where  $w$  is given by (1.1)

**Remarks.** 1. If  $\mu, \nu$  are two probability measures, then

$$S(\nu | \mu) = \begin{cases} - \int \log \left( \frac{d\nu}{d\mu} \right) d\nu, & \text{if } \nu \text{ is } \mu\text{-a.c.} \\ -\infty, & \text{otherwise.} \end{cases} \quad (3.8)$$

is called the relative entropy. Its negative is called the Kullback–Leibler (KL) divergence by statisticians. Both sign conventions are used for “the relative entropy”. We follow much of the spectral theory literature which follows Killip–Simon [26]. Most of the probability literature uses the opposite sign! (3.7) is  $-S(\frac{d\theta}{2\pi} | \mu)$ . That  $I$  is lower semicontinuous goes back to Verblunsky [43]. That the reverse Kullback–Leibler divergence is relevant to large deviations appears in Lynch–Sethuraman [32] and Ganesh–O’Connell [21].

2. It is remarkable (though not unusual) that for each  $n$ ,  $\mathbb{P}_n$  is supported on a set where the rate function (i.e. the Kullback-Leibler) divergence is infinite. This is since the approximating measures are discrete while the limiting one is absolutely continuous.

As a preliminary, one needs to look at what spectral theorists call the density of states, OP workers the density of zeroes and probabilists the empirical measure, namely

$$\mu^{(E)} = \frac{1}{n} \sum_{j=1}^n \delta_{\lambda_j} \quad (3.9)$$

where  $\lambda_j$  are the atoms of  $\mu$  in (3.1).  $\mathbb{P}_n$  induces a distribution  $\mathbb{P}_n^{(E)}$  on point measures of the form (3.9), essentially given by (3.4)/(3.5). One has the following result of Ben Arous and Guionnet [4] (see also [3, Section 2.6]; these results discuss GUE, not CUE – the analog for CUE uses the same ideas and is even simpler):

**Theorem 3.2.**  $\mathbb{P}_n^{(E)}$  obeys a LDP with speed  $n^2$  and good rate function

$$I(\mu) = - \int \log(|z - w|) d\mu(z) d\mu(w) \quad (3.10)$$

**Remark.** In (3.10),  $z$  and  $w$  lie in the unit circle and  $|z - w|$  is a two dimensional distance. This is a  $2D$  Coulomb energy. There is a close connection between this result and Johansson's proof [24] of the Strong Szegő Theorem.

We will not give a formal proof of Theorem 3.2 but instead indicate the basic intuition. For distinct  $\lambda_i$ s,

$$\prod_{i < j} |e^{i\theta_i} - e^{i\theta_j}|^2 = \exp(-n^2 J_n(\lambda_1, \dots, \lambda_n)) \quad (3.11)$$

$$\begin{aligned} J_n(\lambda_1, \dots, \lambda_n) &= -\frac{2}{n^2} \sum_{i < j} \log(|\lambda_i - \lambda_j|) \\ &= -\frac{1}{n^2} \sum_{i \neq j} \log(|\lambda_i - \lambda_j|) \end{aligned} \quad (3.12)$$

If  $\mu^{(E)}$  is an  $n$ -point measure near  $\mu$  and the  $\lambda$  have reasonable local spacing, the sum in (3.12) should be near the integral in (3.10). This completes our description of the intuition behind the proof of Theorem 3.2.

The weights and eigenvalues are independent. We'll consider a fixed triangular array of eigenvalues  $\{\lambda_\ell^{(n)}\}_{1 \leq \ell \leq n; n=1, \dots}$  where we suppose that

$$\frac{1}{n} \sum_{\ell=1}^n \delta_{\lambda_\ell^{(n)}} \rightarrow \frac{d\theta}{2\pi} \quad (3.13)$$

weakly. We distribute weights uniformly on the simplex and look at

$$\{w_\ell\}_{\ell=1}^n \mapsto \sum_{\ell=1}^n w_\ell \delta_{\lambda_\ell^{(n)}} \equiv \mu_n(w_\ell) \quad (3.14)$$

This gives a distribution,  $\mathbb{P}_n^{(\lambda)}$ , on measures and we'll prove these measures obey a LDP with speed  $n$  and rate function  $I$  given by (3.7). A full analysis depends on proving for each  $\epsilon > 0$ ,  $j$  and  $k = 1, \dots, 2^j$ , the probability that  $\left| \frac{2^j}{n} \#(\ell \mid \lambda_\ell^{(n)} \in I_k^{(j)}) - 1 \right| \geq \epsilon$  (with  $I_k^{(j)}$  given by (2.18)) goes to zero faster than exponentially in  $n$ . This is where Theorem 3.2 is used.

The proof will be to use projective limits with the maps of Example 2.9. We'll get a LDP for the projections using Example 2.4 and control the sup of the projected rate functions by a general continuity result. It is this last fact that will show singular parts of the measure only change the rate by their impact on the total weight of the a.c. part.

For each  $j = 1, \dots$  and  $k = 1, \dots, 2^j$ , let  $I_k^{(j)}$  be given by (2.18) and  $\pi_j(\mu)$  the measure in (2.19). Given  $\{w_\ell\}_{\ell=1}^n$ , let  $\tilde{\mu}_n^j(w_\ell)$  be the measure on  $\partial\mathbb{D}$  with constant a.c. weight on each  $I_k^{(j)}$  so that

$$\tilde{\mu}_n^j(I_k^{(j)}) = \sum_{\lambda_\ell^{(n)} \in I_k^{(j)}} w_\ell \quad (3.15)$$

Thus in terms of the objects given in (3.14) and (3.15), we have that  $\pi_j(\mu_n(w_\ell)) = \tilde{\mu}_n^j(w_\ell)$ .

Let  $\tilde{\mathbb{P}}_n^{(j)}$  be the measure on  $\mathbb{R}^{2^j}$  using (3.15) but where now the  $w_\ell$  are replaced by iid exponential random variables,  $W_\ell$ . Thus,  $\tilde{\mathbb{P}}_n^{(j)}$  is the probability measure for the  $\mathbb{R}^{2^j}$ -valued random variable given by

$$\beta_k^n = \sum_{\lambda_\ell^{(n)} \in I_k^{(j)}} W_\ell$$

Fix  $j$  and take  $n \rightarrow \infty$ . By Theorem 2.5 and (3.13),  $\tilde{\mathbb{P}}_n^{(j)}$  obeys a LDP with speed  $n$  and rate function at the point  $\vec{\beta} \equiv \{\beta_\ell\}_{\ell=1}^{2^j} \in \mathbb{R}^{2^j}$

$$\varphi(\vec{\beta}) = \sum_{\ell=1}^{2^j} [(\beta_\ell - 2^{-j}) - 2^{-j} \log(2^j \beta_\ell)] \quad (3.16)$$

Recall that given two probability measures  $\mu$  and  $\nu$  on the same space, their KL divergence,  $H(\mu|\nu)$ , is given by the negative of (3.8). Write  $\beta_\ell = \beta s_\ell$  with  $\beta = \sum_{q=1}^{2^j} \beta_q$  so that  $\vec{s}$  lies in a  $2^j$ -simplex. Write  $\mu_{\vec{s}}$  for the probability measure giving uniform weight  $s_k$  to  $I_k^{(j)}$  and let  $\nu$  be normalized Lebesgue measure on the circle (i.e.  $\mu_{\vec{s}}$  for the  $\vec{s}$  with equal components,  $2^{-j}$ ). Then (3.16) can be rewritten:

$$\varphi(\vec{\beta}) = \beta - 1 - \log(\beta) + H(\nu|\mu_{\vec{s}}) \quad (3.17)$$

Note this is the sum of a function of  $\beta$  only and a function of the  $s$ 's only. This is a consequence of the fact that for independent exponential random variables,  $\sum_{k=1}^N X_k$  is independent of  $\{X_j / \sum_{k=1}^N X_k\}_{j=1}^N$ . It makes the use of the contraction principle (which, in general, is already simple), extremely simple.

For fixed  $\lambda$ 's, let  $\mathbb{P}_n^{(j)} = \pi_j^* \left( \mathbb{P}_n^{(\lambda)} \right)$ . This is just the contraction of  $\tilde{\mathbb{P}}_n^{(j)}$  under the map  $G(\vec{\beta}) \equiv \vec{\beta}/\beta$  from  $\mathbb{R}^{2^j}$  to the  $2^j$ -simplex. By the contraction principle (Theorem 2.7) and

$$\inf_{\beta > 0} [\beta - 1 - \log(\beta)] = 0$$

(as it must as the rate function, (2.12), when  $\alpha = 1$ ), we see that for each fixed  $j$ ,  $\mathbb{P}_n^{(j)}$  obeys a LDP with speed  $n$  and rate function  $H(\nu|\mu_{\vec{s}})$ .

Given the projection theorem (Theorem 2.7), the following completes the proof of Theorem 3.1:

**Proposition 3.3.** *Let  $\mu$  be an arbitrary probability measure on  $\partial\mathbb{D}$  and  $\nu = \frac{d\theta}{2\pi}$ . Let  $\pi_j(\mu)$  be given by (2.19). Then*

$$\lim_{j \rightarrow \infty} H(\pi_j(\nu)|\pi_j(\mu)) = H(\nu|\mu) \quad (3.18)$$

**Remarks.** 1.  $\pi_j(\nu) = \nu$  for this  $\nu$ . We write it this way because with a slight change in the proof, it holds for any  $\nu$  (and  $\mu$ ). This extended version is needed for the Killip–Simon theorem and other cases where the limiting empirical measure is not unweighted Lebesgue measure.

2. By slightly expanding the argument, one sees that  $H(\pi_j(\nu)|\pi_j(\mu))$  is monotone increasing in  $j$ .

3. It is worth repeating that this holds independently of the singular part of  $\mu$  and is responsible for the fact that rate is the same for two  $\mu$ 's with identical a.c. parts independently of their singular parts.

*Proof.* By convexity of  $y \mapsto -\log y$  and Jensen's inequality, for any positive function  $h$  and probability measure  $d\eta(y)$ , we have that

$$-\int \log h(y) d\eta(y) \geq -\log \left( \int h(y) d\eta(y) \right)$$

Thus if  $d\mu$  has the form (1.1) and  $W_k^{(j)} = 2^j \int_{I_k^{(j)}} w(\theta) \frac{d\theta}{2\pi}$ , we have that

$$\begin{aligned} - \int_{I_k^{(j)}} \log(w(\theta)) 2^j \frac{d\theta}{2\pi} &= -\log(W_k^{(j)}) - \int_{I_k^{(j)}} \log\left(\frac{w(\theta)}{W_k^{(j)}}\right) 2^j \frac{d\theta}{2\pi} \\ &\geq -\log W_k^{(j)} \geq -\log\left(2^j \mu(I_k^{(j)})\right) \end{aligned} \quad (3.19)$$

since  $\log\left[\int_{I_k^{(j)}} \frac{w(\theta)}{W_k^{(j)}} 2^j \frac{d\theta}{2\pi}\right] = 0$ .

On the other hand,  $\pi_j(\mu) \equiv \mu^{(j)}$  is an absolutely continuous measure with weight  $\sum_{k=0}^{2^j} 2^j \mu(I_k^{(j)}) \chi_{I_k^{(j)}}(\theta)$  so

$$H(\nu|\mu^{(j)}) = -2^{-j} \sum_{k=0}^{2^j} 2^j \log\left(2^j \mu(I_k^{(j)})\right) \quad (3.20)$$

which by (3.19) implies that

$$H(\pi_j(\nu)|\pi_j(\mu)) \leq H(\nu|\mu) \quad (3.21)$$

so, in particular,

$$\limsup H(\pi_j(\nu)|\pi_j(\mu)) \leq H(\nu|\mu) \quad (3.22)$$

For the other direction,  $\pi_j(\mu) \rightarrow \mu$  weakly by an easy argument (the uniform closure of functions constant on the  $I_k^{(j)}$  for some  $j$  includes all continuous functions). By the lower semicontinuity of  $H$  jointly in the two variables (see [36, Theorem 2.3.4] which proves that  $-H$  is jointly weakly upper semicontinuous)

$$H(\nu|\mu) \leq \liminf H(\pi_j(\nu)|\pi_j(\mu)) \quad (3.23)$$

□

#### 4. SZEGŐ'S THEOREM: COEFFICIENT SIDE

In this section, our goal is to use the bijection  $V$  from Verblunsky coefficients to measures on  $\partial\mathbb{D}$  to move the CUE measures  $\mathbb{P}_n$  from the measure side to measures  $\tilde{\mathbb{P}}_n$  on Verblunsky coefficients. We'll prove:

**Theorem 4.1.**  *$\tilde{\mathbb{P}}_n$  obeys a LDP with speed  $n$  and good rate function*

$$\tilde{I}(\alpha) = - \sum_{j=1}^{\infty} \log(1 - |\alpha_j|^2) \quad (4.1)$$



**Remarks.** 1. Since  $V$  is a homeomorphism, Theorem 3.1 implies that  $\tilde{\mathbb{P}}_n$  obeys a LDP. The point is that (4.1) expresses the rate on the  $\alpha$ -side. In fact, we'll independently prove that  $\tilde{\mathbb{P}}_n$  obeys a LDP. In further applications, it is useful that we only have to prove a LDP on one side.

2. Of course, this result and (3.7) imply the Szegő–Verblunsky sum rule (1.4).

One part of the proof is an explicit formula for  $\tilde{\mathbb{P}}_n$  found by Killip–Nenciu [25]. For a proof, see Killip–Nenciu [25], or the Appendix below.

**Theorem 4.2.**  $\tilde{\mathbb{P}}_n$  is supported on the  $n$ -point  $\alpha$ 's, i.e.  $\alpha_0, \dots, \alpha_{n-2} \in \mathbb{D}, \alpha_{n-1} \in \partial\mathbb{D}$  and given by

$$d\tilde{\mathbb{P}}_n(\alpha_0, \dots, \alpha_{n-1}) = \prod_{j=0}^{n-1} d\kappa_{n-2-j}(\alpha_j) \quad (4.2)$$

$$d\kappa_\ell(\alpha) = \frac{\ell+1}{\pi} (1 - |\alpha|^2)^\ell d^2\alpha \quad \text{on } \mathbb{D}; \ell \geq 0 \quad (4.3)$$

$$d\kappa_{-1}(\alpha = e^{i\theta}) = \frac{d\theta}{2\pi} \quad \text{on } \partial\mathbb{D} \quad (4.4)$$

The density is thus  $\frac{(n-1)!}{\pi^n} \prod_{j=0}^{n-2} (1 - |\alpha_j|^2)^{n-2-j} d^2\alpha_j$ . The  $\prod_{j=0}^{n-2} (1 - |\alpha_j|^2)^n = \exp \left[ -n \left( - \sum_{j=0}^{n-2} \log(1 - |\alpha_j|^2) \right) \right]$  suggests that the rate function is given by (4.1) but naively, the  $(n-1)!$  looks worrying. That it isn't comes from the magic of projective limits.

Projective limits are especially simple in the case of independent variables. So look at measures  $\{\mathbb{P}_N^\#\}_{N=1}^\infty$  on  $X = \mathbb{D}^\infty$  by putting  $\mathbb{P}_N^\#$  on sequences with  $\alpha_j = 0; j \geq N$  and distributing  $\{\alpha_j\}_{j=0}^{N-1}$  according to  $\tilde{\mathbb{P}}_N$ . Then, if  $j$  is fixed and  $N > j$ , then

$$d\pi_j^*(\mathbb{P}_N^\#)(\alpha_0, \dots, \alpha_{j-1}) = \frac{(N-1) \cdots (N-j)}{\pi^j} \left[ \prod_{k=0}^j (1 - |\alpha_k|^2) \right]^{N-2-j} \prod_{k=0}^{j-1} [(1 - |\alpha_k|^2)^{j-k} d^2\alpha_k] \quad (4.5)$$

The leading factor is polynomial in  $N$ , so unimportant for  $a_N = N$  LDPs. Using Theorems 2.1 and 2.2, we see that  $\pi_j^*(\mathbb{P}_N^\#)$  obeys a LDP at speed  $N$  and rate function

$$I_j(\alpha_0, \dots, \alpha_{j-1}) = - \sum_{k=0}^{j-1} \log(1 - |\alpha_k|^2) \quad (4.6)$$

Since

$$\sup_j I_j(\pi_j(\alpha)) = - \sum_{k=0}^{\infty} \log(1 - |\alpha_k|^2) \quad (4.7)$$

this proves, on account of Theorem 2.7, that  $\{\mathbb{P}_N^\#\}_{N=1}^\infty$  obeys a LDP with speed  $N$  and rate function (4.1).

The measures we are really interested in aren't on the product space  $X$  but on the space,  $Y$ , which is the union of  $\mathbb{D}^\infty$  with finite sequences in  $\mathbb{D}^{N-2} \times \partial\mathbb{D}$  with the topology described in Section 1. There is a natural map  $f : X \rightarrow Y$ , that given  $\alpha \in X$  maps to  $\alpha$  if all  $\alpha_j \in \mathbb{D}$  and drops the  $\alpha_k$ ,  $k > j$  is  $\alpha_j$  if the first  $\alpha_\ell \in \partial\mathbb{D}$ . Then  $f$  is continuous and  $(\mathbb{P}_N^\#)^{(f)} = \tilde{\mathbb{P}}_N$ , so Theorem 2.6 completes the proof of Theorem 4.1.

## 5. THE KILLIP–SIMON THEOREM

The large deviation proof of the Killip–Simon Theorem is similar to the one in Sections 3 and 4 with some changes and additions which we briefly describe.

- (a) One uses GUE instead of CUE (GNR use GOE which differs from GUE by some factors of 2). Thus the measure on random  $n \times n$  self-adjoint matrices has  $\{\operatorname{Re} M_{ij}^{(n)}\}_{1 \leq i \leq j \leq n}$  and  $\{\operatorname{Im} M_{ij}^{(n)}\}_{1 \leq i < j \leq n}$  Gaussian iid with mean zero and  $\mathbb{E}([M_{ii}^{(n)}]^2) = \mathbb{E}([\operatorname{Re} M_{ij}^{(n)}]^2) = \mathbb{E}([\operatorname{Im} M_{ij}^{(n)}]^2) = n^{-1}$  for any  $i < j$ .
- (b) The eigenvalue distribution has  $\lambda_j \in \mathbb{R}$  with distribution

$$\left[ \prod_{i < j} |\lambda_i - \lambda_j|^2 \right] e^{-n \sum_{j=1}^n \lambda_j^2} \quad (5.1)$$

so the empirical measure converges to the equilibrium measure in a quadratic external field, i.e. the minimizer for  $-\int \log|x - y| d\mu(x) d\mu(y) + 2 \int x^2 d\mu(x)$ . It is well-known [3, Exercise 2.6.4] that the minimizer is the semicircle law  $d\nu_0(x) \equiv \pi^{-1}(1 - x^2)^{1/2} \chi_{[-1,1]}(x) dx$ . To agree with the Killip–Simon notation, one rescales the matrix so the support is  $[-2, 2]$ .

- (c) By the argument of [4], the empirical measure converges to  $\nu_0$  and by mimicking the proof of Theorem 3.1, the contribution of the part of the spectral measure on  $[-2, 2]$  is just  $H(\nu_0|\mu)$ . Thus the weight in the Killip–Simon quasi–Szegő integral is exactly the Wigner semicircle weight.
- (d) As we've seen, a single point in the measure, if the point is in the bulk, involves the increase of  $H(\nu|\mu)$  due to the weight having a smaller integral. But if the point is outside  $[-2, 2]$ , there is a

contribution due to the location,  $\lambda_0$ , of the eigenvalue (the impact of the weight is the same whether the point is in the bulk or outside it). By looking at the log of the part of (5.1) depending on  $\lambda_0$ , one sees that the decrease in the eigenvalue density involves  $\lambda_0$  interacting with  $n$  eigenvalues. The decrease is approximately  $\exp(-nF(\lambda_0))$  where  $F$  is the potential in the quadratic external field in the equilibrium measure (this idea is due to Ben Arous et al. [5]). It is known that this function is the same as the  $F$  in equation (1.8) (see [1, eqn 1.13] or [11, proof of Thm 3.6]) so the Killip–Simon  $F$  is just an external field potential.

- (e) For finitely many eigenvalues outside  $[-2, 2]$  you just get the sums of single costs since the interaction between eigenvalues is  $O(1)$ , not  $O(n)$ . Handling infinitely many eigenvalues converging to  $\pm 2$  requires a careful use of projective limits (see [17]).
- (f) For the coefficient side, Killip–Nenciu is replaced by earlier results of Dumitriu–Edelman [16] (whose work motivated Killip and Nenciu) who found the distribution of Jacobi parameters for GUE and GOE. The  $\{b_j\}_{j=1}^n$  are Gaussian (with  $O(n)$  widths leading to the  $b_j^2$  term in the Killip–Simon sum rule). The  $\{a_j^2\}_{j=1}^{n-1}$  are gamma distributed, essentially behaving like sums of exponential random variables and so we get the  $G(a_j)$  terms. Thus  $G$  occurs in the sum rule as the rate function for suitable gamma distributions.
- (g) There is an issue involving the equality of the two sides of the sum rule that we want to discuss, addressed in a related way in Gamboa-Rouault [20]. The natural setting for the LDP for measures is the space,  $X'$ , of all probability measures on  $\mathbb{R}$ , and for Jacobi parameters the Polish space  $Y' \equiv [\mathbb{R} \times (0, \infty)]^\infty$  with finite sequences added to it. The issue is that the inverse Jacobi map isn't defined for all measures but only those with all moments finite and, in general, this inverse map is many-to-one in certain cases where the measure has unbounded support. Let  $X_k$  be the set of measures supported in  $[-k-2, k+2]$  for  $k = 1, 2, \dots$  and  $Y_k$  its image under the inverse Jacobi maps. Let  $X = \cup_{k=1}^\infty X_k$  and  $Y = \cup_{k=1}^\infty Y_k$ . The rate functions are infinite on the complements of these sets. The Jacobi map is a well defined bijection of  $Y$  to  $X$  but in the relative topologies is not continuous nor is its inverse continuous! However, it is a homeomorphism restricted to each  $Y_k$ . Moreover since the probabilities under  $\mathbb{P}_N$  (resp.  $\tilde{\mathbb{P}}_N$ ) of  $X_k$  (resp.  $Y_k$ ) go to one exponentially fast in  $n$  (with rate going to infinity with  $k$ ), it is not hard to prove LDP's for the restrictions of these measures to  $X_k$  and  $Y_k$  renormalized (by dividing by the

probabilities of  $X_k$  and  $Y_k$ ) with the same rate functions. Thus the fact that  $J$  is a homeomorphism of  $Y_k$  to  $X_k$  lets us conclude equalities of the two rate functions and so the sum rule.

## 6. FURTHER DEVELOPMENTS

Finally, we want to make a few comments about the strategy of the last three sections applied to other sum rules. We are aware of four papers on this approach. Gamboa, Nagel and Rouault have two preprints [18, 19] besides other results in their original paper [17]. And the authors of the current paper are preparing a work [7] using these methods in a wider context.

In [36, Section 2.8], Simon found a sum rule involving  $-\int(1 - \cos(\theta)) \log(w(\theta)) \frac{d\theta}{2\pi}$  on the measure side and made a conjecture concerning

$$-\int \log(w(\theta)) d\eta(\theta) \quad (6.1)$$

where

$$d\eta(\theta) = Z^{-1} \prod_{j=1}^k (1 - \cos(\theta - \theta_j))^{m_j} d\theta \quad (6.2)$$

where  $Z$  is a normalization factor to make  $d\eta$  into a probability measure. There developed a huge literature on these so called higher order sum rules for OPUC and OPRL including [13, 22, 28, 29, 30, 31, 33].

The key to understanding such sum rules (for OPUC) in the context of large deviations is to replace Haar measure,  $d\mathbb{P}_N$ , by

$$Z_N^{-1} \exp \left[ -N \sum_{j=1}^N V(\lambda_j) \right] d\mathbb{P}_N \quad (6.3)$$

where  $V$  is a function on  $\partial\mathbb{D}$  and  $\{\lambda_j\}_{j=1}^N$  are the eigenvalues. It is well known (see [3, Section 2.6]) that when  $V$  is nice enough, we will get  $d\eta$  as the equilibrium measure if

$$V(e^{i\theta}) = 2 \int \log |e^{i\theta} - e^{i\psi}| d\eta(\psi) \quad (6.4)$$

In a forthcoming paper [7], the current authors study this when  $d\eta$  is given by (6.2). In the cases we study,  $V(e^{i\theta})$  is a finite linear combination of  $\cos(m\theta)$ . In terms of  $U$ , if  $e^{i\theta_j}$  are the eigenvalues,  $\sum_{j=1}^n \cos(m\theta_j) = \operatorname{Re}(\operatorname{Tr}(U^m))$  which one can write in terms of Verblunsky coefficients using the CMV representation of  $U$ . We obtain a large deviations proof of the  $(1 - \cos(\theta))$  sum rule of Simon and the gems of Simon–Zlatoš [42]. In addition, we prove a partial special case of a conjecture of Lukic [30] that replaces a wrong conjecture of Simon [36,

Section 2.8], providing evidence for Lukic's conjecture. GNR have a paper [19] that discusses in some detail the case  $V(\theta) = \cos(\theta)$  where the random matrix model has been studied by Gross–Witten [23] whose names GNR apply to the model. They note that formally the large deviations argument leads to a sum rule but for technical reasons, they aren't able to provide a proof. By using some results from the theory of OPUC, we do prove sum rules in this and the other cases.

In their original paper [17], GNR introduce two new results they call magic sum rules by using large deviations on two matrix models. These models have free parameters and lead to families of sum rules. In these models the continuous spectrum is an interval  $[\alpha, \beta]$ . There are three classes of sum rules where the KL divergence has  $H(\nu|\mu)$  with

$$d\nu(x) = \chi_{[\alpha, \beta]}(x)(x - \alpha)^{\sigma/2}(\beta - x)^{\tau/2} dx$$

with  $(\sigma, \tau)$ , one of  $(1, 1)$ ,  $(-1, -1)$  or  $(-1, 1)$  (or  $(1, -1)$ ). All these new sum rules lead to what one of us calls flawed gems because they have apriori conditions on the Jacobi parameters. This is because the sum rules only hold under these conditions (the restriction comes from the fact that the Jacobi parameters have to be expressible in terms of other more basic parameters of the underlying model and only certain Jacobi parameters can be expressed that way). The  $(1, 1)$  examples lead to gems that are restricted forms of the Killip–Simon Theorem and so not new gems. The  $(-1, -1)$  examples are restricted forms of Szegő's Theorem under the Szegő maps (see [37, Section 13.1]) and so are not new. Their  $(-1, 1)$  examples yield new flawed gems.

There has been considerable literature on proving analogs of the Killip–Simon theorem where  $[-2, 2]$  is replaced by a finite gap set  $\mathfrak{e} = \mathfrak{e}_1 \cup \dots \cup \mathfrak{e}_n$  where the  $\mathfrak{e}_j$  are disjoint intervals. In [9], Damanik et al prove a Killip–Simon rule in the case that each  $\mathfrak{e}_j$  has harmonic measure (i.e. measure in the equilibrium measure for  $\mathfrak{e}$ )  $1/n$ . They first prove a Killip–Simon sum rule for  $[-2, 2]$  but with matrix valued measures and then use something they call the magic formula to get a gem for these special sets. In [18], GNR use large deviations to get a new proof of the Killip–Simon sum rule on  $[-2, 2]$  with matrix valued measures and then plug that into the DKS magic formula machine to get a partially new proof of the DKS result for  $1/n$  harmonic measure sets. It would be very interesting to obtain this result directly with large deviations without using the magic formula.

Recently Yuditskii [44] proved an analog of the Killip–Simon Theorem for any finite gap set,  $\mathfrak{e}$ . It would be very interesting to find a large deviations proof of his result.

There has been very little work on Killip–Simon type theorems for finite gap sets in OPUC. In [19], GNR obtain a sum rule and gem for  $\mathfrak{e} = \{e^{i\theta} \mid \alpha \leq \theta \leq 2\pi - \alpha\}$  for  $0 < \alpha < \pi$ . For real  $\alpha$ , the Verblunsky side has the expected  $\sum |\alpha_j - a|^2$  form but for general  $\alpha$ , it has the form  $\sum |\gamma_j - a|^2$  where  $\gamma_j$  is a non-local function of the  $\alpha$ 's. In particular, it is not clear if the finiteness of their Verblunsky side only depends on the behavior near  $j = \infty$ . At least for the real case, it would be interesting to get the sum rule via the Poisson–Jensen methods used in the original Killip–Simon proof [35]. It would also be interesting to understand the  $\gamma_j$ 's in a more conventional setting.

Finally, we note that Killip–Simon [27] have proven a sum rule and gem for half-line Schrödinger operators when  $V \in L^2((0, \infty); dx)$ . It would be very interesting to find a large deviation proof of this result. In particular, what is the analog of random matrix models for the study of Schrödinger operators.

#### APPENDIX A.

We want to describe a proof of Theorem 4.2 that, because it is inductive, may be attractive to the reader. We note that if one uses the GGT proof of the critical lemma below and uses explicit iteration instead of induction, our proof translates into a variant of the original proof of [25]. We begin by writing  $\mathbb{U}(n)$ , the group of  $n \times n$  unitary matrices, as a product of  $\mathbb{U}(n-1)$  and  $\mathbb{C}_1^n$ , the set of unit vectors in  $\mathbb{C}$ . Since topologically  $\mathbb{U}(n)$  is not the product of  $\mathbb{U}(n-1)$  and  $\mathbb{S}^{2n-1}$ , the unit sphere in  $\mathbb{R}^{2n}$ , this association cannot be continuous, but it will be measurable. Critically, Haar measure on  $\mathbb{U}(n)$  will be a product measure of Haar measure on  $\mathbb{U}(n-1)$  and the rotation invariant measure on  $\mathbb{C}_1^n$ .

This idea goes back at least to Diaconis-Shahshahani [15] who discussed the relation of Haar measure on  $G$  to the natural product measure on  $H \times G/H$  in general and illustrated this for the orthogonal group,  $\mathbb{O}(n)$ , using the Householder algorithm. The discussion below is just the  $\mathbb{U}(n)$ -analog of their discussion of  $\mathbb{O}(n)$ . To be specific, let  $G$  be a compact group and  $H$  a closed subgroup of  $G$ . Let  $\pi : G \rightarrow G/H$  be the canonical projection. Normalized Haar measure,  $\mu_G$ , induces a natural probability measure,  $\mu_{G/H}$ , on  $G/H$  via

$$\mu_{G/H}(A) = \mu_G(\pi^{-1}[A]) \quad (\text{A.1})$$

and this measure is clearly invariant under the action of  $G$  on  $G/H$ .

Let  $\sigma : G/H \rightarrow G$  be a choice of representative from each coset, i.e.  $\pi(\sigma(x)) = x$ . Then  $\Sigma : G/H \times H \rightarrow G$ , defined by  $\Sigma(x, h) = \sigma(x)h$ , is a bijection. If one can choose  $\sigma$  to be continuous, then  $G$  will be

homeomorphic to  $G/H \times H$  under  $\Sigma$  and often such a homeomorphism doesn't exist, e.g. if  $G = \mathbb{U}(n)$  and  $H = \mathbb{U}(n-1)$ , so we should avoid the assumption that  $\sigma$  is continuous. It is probably true that in general one can make a measurable choice. Since we'll find an explicit such choice below for the case of interest we shall simply suppose that  $\sigma$  is measurable.

**Proposition A.1** (Diaconis-Shahshahani [15]). *Suppose  $\sigma$  is measurable. Then under the bijection  $\Sigma$  of  $G/H \times H$  and  $G$ , the measure  $\mu_{G/H} \otimes \mu_H$  goes to  $\mu_G$ .*

*Proof.* Let  $U \in G$ ,  $x \in G/H$ . Then  $\pi(U\sigma(x)) = Ux$  so for some  $W_{U,x} \in H$ , we have that

$$U\sigma(x) = \sigma(Ux)W_{U,x} \quad (\text{A.2})$$

so  $U\Sigma(x, W) = \Sigma(Ux, W_{U,x}W)$  which, given the invariance of  $\mu_{G/H}$  under the action of  $G$  and of  $\mu_H$  under left multiplication by elements of  $H$ , implies the image of the product measure is invariant under multiplication by  $U$  (by integrating first over  $W$  and then  $x$ ).  $\square$

Returning to  $\mathbb{U}(n)$ , fix a unit vector  $e_1 \in \mathbb{C}_1^n$  (it may be helpful to think of  $e_1$  as the first vector,  $\delta_1 = (1, 0, \dots, 0)$ , of the canonical basis of  $\mathbb{C}^n$ ). The map  $U \mapsto Ue_1$  is a surjection of  $\mathbb{U}(n)$  to  $\mathbb{C}_1^n$ . The inverse image of  $e_1$  is those unitaries of the form  $U = \mathbf{1} \oplus W$ , under the direct sum decomposition  $\mathbb{C}^n = [e_1] \oplus [e_1]^\perp$ , where  $W$  is an arbitrary unitary on  $[e_1]^\perp$ . Thus the set of  $W$ 's is isomorphic to  $\mathbb{U}(n-1)$  and, if  $e_1 = \delta_1$ , is canonically equal to it. This shows that the quotient group  $\mathbb{U}(n)/\mathbb{U}(n-1)$  of left cosets of  $\mathbb{U}(n-1)$  is exactly  $\mathbb{C}_1^n$ .

To realize  $\mathbb{U}(n)$  as a product of  $\mathbb{U}(n-1)$  and  $\mathbb{C}_1^n$ , we must pick, for each  $f \in \mathbb{C}_1^n$ , an element  $\sigma(f) \in \mathbb{U}(n)$  so that  $\sigma(f)e_1 = f$ , i.e.  $\sigma(f)$  is in the coset associated to  $f$ . By the above noted fact about topological products, we cannot make this choice continuous in  $f$ , but one can make it measurable, indeed continuous on  $\mathbb{C}_1^n \setminus \{\mathbb{C} \cdot e_1\}$ , as follows. Suppose  $f$  is not colinear with  $e_1$ . Then  $e_1, f$  span a two dimensional subspace  $\mathcal{H}_f$ . We can pick another vector  $e_2 \in \mathcal{H}_f$  orthonormal to  $e_1$  specifying it uniquely by demanding that

$$\kappa \equiv \langle f, e_2 \rangle > 0 \quad (\text{A.3})$$

We also define

$$\beta \equiv \overline{\langle f, e_1 \rangle} \quad (\text{A.4})$$

so that  $\beta \in \mathbb{D}$  and

$$f = \bar{\beta}e_1 + \kappa e_2 \quad (\text{A.5})$$

Since  $f$  is a unit vector

$$\kappa = \sqrt{1 - |\beta|^2} \quad (\text{A.6})$$

There is an obvious unitary map on  $\mathcal{H}_f$  that takes  $e_1$  to  $f$ , namely reflection,  $\Theta(\beta)$ , in the line along  $e_1 + f$ , which is clearly

$$\mathbf{1} - 2\langle e_1 - f, \cdot \rangle (e_1 - f) / \|e_1 - f\|^2 \quad (\text{A.7})$$

To find its matrix form in the  $e_1, e_2$  basis, we note by (A.5), that its first column must be  $\begin{pmatrix} \bar{\beta} \\ \kappa \end{pmatrix}$  since it takes  $e_1$  to  $f$ . Its second column is then determined by orthonormality and the desire to have determinant -1 (i.e. a reflection). Thus

$$\Theta(\beta) = \begin{pmatrix} \bar{\beta} & \kappa \\ \kappa & -\beta \end{pmatrix} \quad (\text{A.8})$$

We define the Householder reflection,  $\sigma(f)$ , on  $\mathbb{C}^n$  to be  $\Theta(\beta) \oplus \mathbf{1}_{n-2}$  under  $\mathbb{C}^n = \mathcal{H}_f \oplus \mathcal{H}_f^\perp$  (where  $\mathbf{1}_k$  is the size  $k$  identity matrix).  $\sigma(f)$  is given by (A.7), now as an operator on  $\mathbb{C}^n$ . This formula makes it clear that  $f \mapsto \sigma(f)$  is continuous on  $\mathbb{C}^n \setminus \{\mathbb{C} \cdot e_1\}$  and discontinuous at the points of  $\mathbb{C} \cdot e_1$ . We define  $\sigma$  at  $\lambda e_1$  to be  $\lambda \mathbf{1}$ . We thus have:

**Proposition A.2** (Diaconis-Shahshahani [15]). *Every unitary  $U \in \mathbb{U}(n)$  can be uniquely written  $\sigma(f)W$  where  $f = Ue_1$  and  $W$  is a unitary map on  $[e_1]^\perp$ . This maps  $\mathbb{U}(n)$  Borel bijectively to  $\mathbb{C}_1^n \times \mathbb{U}(n-1)$ . Under this map, Haar measure on  $\mathbb{U}(n)$  is just the product of the rotation invariant measure on  $\mathbb{C}_1^n \cong \mathbb{S}^{2n-1}$  and Haar measure on  $\mathbb{U}(n-1)$ .*

The final assertion follows from Proposition A.1.

To link this to OPUC, assume that  $e_1$  is cyclic for  $U$  and let  $d\mu$  be the spectral measure associated to  $(U, e_1)$ . By the spectral theorem, there is a unitary transformation  $V : \mathbb{C}^n \rightarrow L^2(\partial\mathbb{D}, d\mu)$  such that  $Ve_1 = 1$  (the constant function) and  $(VUV^{-1}h)z = zh(z)$ , so that  $(VUe_1)(z) = z$ . In terms of the orthogonal polynomials w.r.t.  $\mu$ , this means that  $Ve_1 = \Phi_0 = \frac{\Phi_0}{\|\Phi_0\|} =: \varphi_0$  and  $Ve_2 = \frac{\Phi_1}{\|\Phi_1\|} =: \varphi_1$  (the orthonormal polynomials). Szegő recursion (1.2) for the normalized polynomials says (with  $\rho_0 = \sqrt{1 - |\alpha_0|^2}$ )

$$z\varphi_0 = \rho_0\varphi_1 + \bar{\alpha}_0\varphi_0 \quad (\text{A.9})$$

so comparing with (A.5), we see that  $\beta = \alpha_0$  and  $\kappa = \rho_0$  and

$$\Theta(\alpha) = \begin{pmatrix} \bar{\alpha} & \rho \\ \rho & -\alpha \end{pmatrix} \quad (\text{A.10})$$

The key lemma is



**Lemma A.3.** *Let  $e_1$  and  $e_2$  be two orthonormal vectors in  $\mathbb{C}_1^n$ . Let  $U'$  be a unitary on  $[e_1]^\perp$  and  $\Theta(\tilde{\alpha})$  given by (A.10) in  $e_1, e_2$  basis. Let*

$$U = [\Theta(\tilde{\alpha}) \oplus \mathbf{1}_{n-2}][\mathbf{1}_1 \oplus U'] \quad (\text{A.11})$$

*Then  $e_1$  is cyclic for  $U$  if and only if  $e_2$  is cyclic for  $U'$ . If they are cyclic, the Verblunsky coefficients,  $\{\alpha_j\}_{j=0}^{n-1}$  for  $(U, e_1)$  and  $\{\beta_j\}_{j=0}^{n-2}$  for  $(U', e_2)$  are related by*

$$\alpha_0 = \tilde{\alpha}; \quad \alpha_j = \beta_{j-1}, \quad j = 1, \dots, n-1 \quad (\text{A.12})$$

**Remark.** The proof will rely on matrix realizations of multiplication by  $z$  on  $L^2(\partial\mathbb{D}, d\mu)$ , the subject of [36, Chapter 4]. There are three such representations there. First the GGT matrix,  $\mathcal{G}$ , which uses the orthonormal basis of  $\mathbb{C}^n$  (assuming  $\mu$  is an  $n$ -point measure) obtained by applying Gram-Schmidt to  $1, z, z^2, \dots$  (i.e. the orthonormal polynomials). Second the CMV matrix,  $\mathcal{C}$ , in the basis obtained by using Gram-Schmidt on  $1, z, z^{-1}, z^2, z^{-2}, \dots$ , and finally the alternate CMV matrix,  $\tilde{\mathcal{C}}$ , obtained using the basis obtained by applying Gram-Schmidt to  $1, z^{-1}, z, z^{-2}, z^2, \dots$ . Below we'll give a proof exploiting  $\mathcal{C}$  and  $\tilde{\mathcal{C}}$  and afterwards indicate a proof using  $\mathcal{G}$ .

*Proof.* The cyclicity statement is a simple calculation.

We recall the  $\mathcal{LM}$  factorization of the CMV matrix [36, Section 4.2]. Define  $\Theta_{-1}$  to be the  $1 \times 1$  identity matrix and, if  $|\alpha| = 1$ ,  $\Theta(\alpha)$  is the  $1 \times 1$  matrix with element  $\bar{\alpha}$ . If  $\alpha \in \mathbb{D}$ ,  $\Theta(\alpha)$  is given by (A.10). If  $n$  is even, we define as operators on  $\mathbb{C}^n$ :

$$\mathcal{L} = \Theta(\alpha_0) \oplus \Theta(\alpha_2) \oplus \dots \oplus \Theta(\alpha_{n-2}); \quad \mathcal{M} = \Theta_{-1} \oplus \Theta(\alpha_1) \oplus \dots \oplus \Theta(\alpha_{n-1}) \quad (\text{A.13})$$

and if  $n$  is odd

$$\mathcal{L} = \Theta(\alpha_0) \oplus \Theta(\alpha_2) \oplus \dots \oplus \Theta(\alpha_{n-1}); \quad \mathcal{M} = \Theta_{-1} \oplus \Theta(\alpha_1) \oplus \dots \oplus \Theta(\alpha_{n-2}) \quad (\text{A.14})$$

If  $\mu$  is an  $n$ -point measure, let  $\{\chi_j\}_{j=0}^{n-1}$  be the orthonormal basis obtained using Gram-Schmidt on  $1, z, z^{-1}, \dots, z^k$  if  $n = 2k$  and  $1, z, z^{-1}, \dots, z^k, z^{-k}$  if  $n = 2k + 1$ . Let  $\{x_j\}_{j=0}^{n-1}$  be the same for  $1, z^{-1}, z, \dots, z^{-k}$  if  $n = 2k$  or  $1, z^{-1}, z, \dots, z^{-k}, z^k$  if  $n = 2k + 1$ . One defines the CMV and alternate CMV matrices by

$$\mathcal{C}_{k\ell} = \langle \chi_k, z\chi_\ell \rangle, \quad \tilde{\mathcal{C}}_{k\ell} = \langle x_k, zx_\ell \rangle \quad (\text{A.15})$$

Then

$$\mathcal{L}_{k\ell} = \langle \chi_k, zx_\ell \rangle, \quad \mathcal{M}_{k\ell} = \langle x_k, \chi_\ell \rangle \quad (\text{A.16})$$

so we have that

$$\mathcal{C} = \mathcal{LM}; \quad \tilde{\mathcal{C}} = \mathcal{ML} \quad (\text{A.17})$$

Clearly

$$\mathcal{L}(\alpha_0, \alpha_2, \dots) = [\Theta(\alpha_0) \oplus \mathbf{1}_{n-2}][\mathbf{1}_2 \oplus \mathcal{L}(\alpha_2, \dots)] \quad (\text{A.18})$$

$$= [\Theta(\alpha_0) \oplus \mathbf{1}_{n-2}][\mathbf{1}_1 \oplus \mathcal{M}(\alpha_2, \dots)] \quad (\text{A.19})$$

Clearly,

$$\mathcal{M}(\alpha_1, \alpha_3, \dots) = \mathbf{1}_1 \oplus \mathcal{L}(\alpha_1, \alpha_3, \dots) \quad (\text{A.20})$$

and this, combined with (A.19), implies the critical

$$\mathcal{C}(\alpha_0, \alpha_1, \dots) = [\Theta(\alpha_0) \oplus \mathbf{1}_{n-2}][\mathbf{1}_1 \oplus \tilde{\mathcal{C}}(\alpha_1, \alpha_2, \dots)] \quad (\text{A.21})$$

One consequence of (A.21) is that the map  $(\alpha_0, \alpha_1, \dots) \mapsto \mathcal{C}(\alpha_0, \alpha_1, \dots)$  is injective, a critical part of a spectral theoretic proof of Verblunsky's Theorem ([36, Theorem 4.2.8] has a more awkward proof of this fact), for  $\alpha_0 = \langle \delta_1, \mathcal{C}(\alpha_0, \alpha_1, \dots) \delta_1 \rangle$ ,  $\alpha_1 = \langle \delta_2, (\Theta(\alpha_0) \oplus \mathbf{1}_{n-2})^{-1} \mathcal{C}(\alpha_0, \alpha_1, \dots) \delta_2 \rangle$ ,  $\alpha_2 = \langle \delta_3, (\Theta(\alpha_0) \oplus \mathbf{1}_{n-2})^{-1} \mathcal{C}(\alpha_0, \alpha_1, \dots) (\mathbf{1}_1 \oplus \Theta(\alpha_1) \oplus \mathbf{1}_{n-3})^{-1} \delta_3 \rangle$ , ....

Moreover, this inductive argument shows the initial basis is the CMV basis for the matrix. A similar set of arguments work for the alternate CMV matrix and its basis

Another consequence of (A.21) is the proof of (A.12). For given  $U, U', e_1, e_2$  as in the hypothesis,  $e_1$  and  $e_2$  are the first two elements in the CMV basis for  $(U, e_1)$ . Extend  $e_1, e_2$  to the full CMV basis for  $(U, e_1)$  (call its elements  $\{e_1, e_2, e_3, \dots, e_n\}$ ). In this basis,  $U = \mathcal{C}(\alpha_0, \alpha_1, \dots, \alpha_{n-1})$ , so (A.11) and (A.21) imply that  $U' = \mathbf{1}_1 \oplus \tilde{\mathcal{C}}(\alpha_0, \alpha_1, \dots, \alpha_{n-1})$ . By the remark at the end of the previous paragraph and (A.21),  $\{e_2, e_3, \dots, e_n\}$  is the alternate CMV basis for  $\tilde{\mathcal{C}}$  and so, since  $\tilde{\mathcal{C}}$  determines its  $\alpha$ 's, we conclude (A.12).  $\square$

**Remark.** For GGT matrices, what Simon [38] calls the AGR factorization after Ammar, Gragg and Reichel [2], implies that

$$\mathcal{G}(\alpha_0, \alpha_1, \dots, \alpha_{n-1}) = [\Theta(\alpha_0) \oplus \mathbf{1}_{n-2}][\mathbf{1}_1 \oplus \mathcal{G}(\alpha_1, \dots, \alpha_{n-1})] \quad (\text{A.22})$$

This can replace (A.21) in an alternate proof of the Lemma and this alternate proof is related to the argument in [25]. Moreover, Simon [38] has a version of the critical Lemma A.3 proven using the AGR factorization.

This also provides new insights into the difference between the GGT and CMV representations. For  $n$  even, define  $\tilde{\mathcal{L}}_j$ ,  $j = 0, 2, \dots, n$  as for  $\mathcal{L}$  with  $\Theta_0, \dots, \Theta_{j-2}, \Theta_{j+2}, \dots, \Theta_{n-2}$  replaced by zero (only  $\Theta_j$  remains in the direct sum) and similarly for  $\tilde{\mathcal{M}}_j$ ,  $j = -1, 1, \dots, n-1$  (where  $\Theta_{-1}, \Theta_{n-1}$  are  $1 \times 1$  matrices. Thus we have that

$$\mathcal{L} = \tilde{\mathcal{L}}_0 + \tilde{\mathcal{L}}_2 + \dots + \tilde{\mathcal{L}}_{n-2}; \quad \mathcal{M} = \tilde{\mathcal{M}}_{-1} + \dots + \tilde{\mathcal{M}}_{n-1} \quad (\text{A.23})$$

Then the  $\tilde{\mathcal{L}}$ 's and  $\tilde{\mathcal{M}}$ 's are all Householder reflections and

$$\mathcal{C} = \tilde{\mathcal{L}}_0 \tilde{\mathcal{L}}_2 \tilde{\mathcal{L}}_{n-2} \dots \tilde{\mathcal{M}}_1 \tilde{\mathcal{M}}_3, \dots \tilde{\mathcal{M}}_{n-1}; \quad \mathcal{G} = \tilde{\mathcal{L}}_0 \tilde{\mathcal{M}}_1 \tilde{\mathcal{L}}_2 \tilde{\mathcal{M}}_3 \dots \tilde{\mathcal{M}}_{n-1} \quad (\text{A.24})$$

with similar formulae if  $n$  is odd.

*Proof of Theorem 4.2.* Fix a vector  $e_1 \in \mathbb{C}_1^n$ . For  $U$  picked according to Haar measure,  $e_1$  is cyclic with probability one. Applying Proposition A.2,  $\tilde{\mathbb{P}}_n(\alpha_0, \dots, \alpha_{n-1})$  is a product measure where  $\alpha_0 = \overline{(e_1, Ue_1)}$  is distributed according to the distribution of  $z_1$  if  $\mathbf{z}$  is uniformly distributed on  $\mathbb{C}_1^n$  which is exactly  $d\kappa_{n-2}$  (since  $\frac{n-1}{\pi}(1 - |v_1|^2)^{n-2}$  is the size of the “slice”  $\{(v_2, \dots, v_n) \in \mathbb{C}^n \mid |(v_2, \dots, v_n)|^2 = (1 - |v_1|^2)\}$ ).

The other factor is what Haar measure for  $\mathbb{U}(n-1)$  induces on the Verblunsky coefficients,  $\beta_0, \dots, \beta_{n-2}$ , for  $\sigma(f)^{-1}U$ . By Lemma A.3, the  $\beta$ 's are given by (A.12), so the result follows by induction.  $\square$

## REFERENCES

- [1] S. Alberverio, L. Pastur, and M. Shcherbina, *On the  $1/n$  expansion for some unitary invariant ensembles of random matrices*, Comm. Math. Phys. **224** (2001) 271–305.
- [2] G. Ammar, W. Gragg, L. Reichel, *On the eigenproblem for orthogonal matrices*, Proceedings of the 25th Conference on Decision and Control, Athens, 1986, pp. 1963–1966.
- [3] G. Anderson, A. Guionnet and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge University Press, 2010.
- [4] G. Ben Arous and A. Guionnet, *Large deviations for Wigner’s law and Voiculescu’s non-commutative entropy*, Probab. Theory Rel. Fields, **108** (1997), 517–542.
- [5] G. Ben Arous, A. Dembo, and A. Guionnet, *Aging of spherical spin glasses*, Probab. Theory Rel. Fields **120** (2001), 1–67.
- [6] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, Inc., New York, 1999.
- [7] J. Breuer, B. Simon and O. Zeitouni, *Large Deviations and the Lukic Conjecture*, in preparation.
- [8] K. Case, *Orthogonal polynomials from the viewpoint of scattering theory*, J. Math. Phys. **15** (1974), 2166–2174.
- [9] D. Damanik, R. Killip and B. Simon, *Perturbations of orthogonal polynomials with periodic recursion coefficients*, Ann. Math. **171** (2010), 1931–2010.
- [10] D. Dawson and J. Gärtner, *Large deviations from the McKean–Vlasov limit for weakly interacting diffusions*, Stochastics, **20** (1987), 247–308.
- [11] P. Deift, T. Kriecherbauer, K. McLaughlin, S. Venakides, and X. Zhou, *Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory*, Comm. Pure Appl. Math. **52** (1999), 1335–1425.
- [12] A. Dembo and O. Zeitouni *Large Deviations and Applications*, 2nd edition, Springer, Berlin, 1998.

- [13] S. Denisov and S. Kupin, *Asymptotics of the orthogonal polynomials for the Szegő class with a polynomial weight*, J. Approx. Theory **139** (2006), 8–28.
- [14] J. Deuschel and D. Stroock, *Large Deviations*, Academic Press, Boston, 1989.
- [15] P. Diaconis and M. Shahshahani, *The Subgroup Algorithm for Generating Uniform Random Variables*, Prob. Eng. Inf. Sc. **1** (1987), 15–32.
- [16] I. Dumitriu, and A. Edelman, A. (2002). *Matrix models for beta ensembles*, J. Math. Phys. **43** (2002), 5830–5847.
- [17] F. Gamboa, J. Nagel, and A. Rouault, *Sum rules via large deviations*, J. Funct. Anal. **270**, (2016), 509–559.
- [18] F. Gamboa, J. Nagel, and A. Rouault, *Sum rules and large deviations for spectral matrix measures*, Preprint. arXiv:1604.06934.
- [19] F. Gamboa, J. Nagel, and A. Rouault, *Sum rules and large deviations for spectral measures on the unit circle*, Preprint. arXiv 1601.08135.
- [20] F. Gamboa and A. Rouault, *Large Deviations for Random Spectral Measures and Sum Rules*, Appl. Math. Res. Express **2** (2011), 281–307.
- [21] A. Ganesh, and N. O’Connell, *A large-deviation principle for Dirichlet posteriors*, Bernoulli **6** (2000), 1021–1034.
- [22] L. Golinskii and A. Zlatoš, *Coefficients of orthogonal polynomials on the unit circle and higher-order Szegő theorems*, Constr. Approx. **26** (2007), 361–382.
- [23] D. Gross and E. Witten, *Possible third-order phase transition in the large- $N$  lattice gauge theory*, Phys. Rev. D **21** (1980), 446–453.
- [24] K. Johansson, *On Szegő’s asymptotic formula for Toeplitz determinants and generalizations*, Bull. Sci. Math. **112** (1988), 257–304.
- [25] R. Killip and I. Nenciu, *Matrix models for circular ensembles*, Int. Math. Res. Not., **50** (2004), 2665–2701.
- [26] R. Killip and B. Simon, *Sum rules for Jacobi matrices and their applications to spectral theory*, Ann. Math. **158** (2003), 253–321.
- [27] R. Killip and B. Simon *Sum rules and spectral measures of Schrödinger operators with  $L^2$  potentials*, Ann. Math. **170** (2009), 739–782.
- [28] S. Kupin, *On a spectral property of Jacobi matrices*, Proc. Amer. Math. Soc. **132** (2004), 1377–1383.
- [29] A. Laptev, S. Naboko and O. Safronov, *On new relations between spectral properties of Jacobi matrices and their coefficients*, Comm. Math. Phys. **241** (2003), 91–110.
- [30] M. Lukic, *On a conjecture for higher-order Szego theorems*, Constr. Approx. **38** (2013), 161–169.
- [31] M. Lukic, *On higher-order Szego theorems with a single critical point of arbitrary order*, Const. Approx., to appear.
- [32] J. Lynch and J. Sethuraman *Large Deviations for Processes with Independent Increments*, Ann. Probab. **15** (1987), 610–627.
- [33] F. Nazarov, F. Peherstorfer, A. Volberg, and P. Yuditskii, *On generalized sum rules for Jacobi matrices*, Int. Math. Res. Not. **3** (2005), 155–186.
- [34] B. Simon *Representations of Finite and Compact Groups*, American Mathematical Society, Providence, 1996.
- [35] B. Simon, *A canonical factorization for meromorphic Herglotz functions on the unit disk and sum rules for Jacobi matrices*, J. Funct. Anal. **214** (2004), 396–409.

- [36] B. Simon, *Orthogonal Polynomials on the Unit Circle, Part 1: Classical Theory*, American Mathematical Society, Providence, RI, 2005.
- [37] B. Simon, *Orthogonal Polynomials on the Unit Circle, Part 2: Spectral Theory*, American Mathematical Society, Providence, RI, 2005.
- [38] B. Simon, *CMV matrices: Five years after*, J. Comput. Appl. Math. **208** (2007), 120–154.
- [39] B. Simon *Szegő's Theorem and Its Descendants: Spectral Theory for  $L^2$  Perturbations of Orthogonal Polynomials*, Princeton University Press, Princeton, NJ, 2011.
- [40] B. Simon, *A Comprehensive Course in Analysis, Part 1, Real Analysis*, American Mathematical Society, Providence, R.I., 2015.
- [41] B. Simon, *A Comprehensive Course in Analysis, Part 2A, Basic Complex Analysis*, American Mathematical Society, Providence, R.I., 2015.
- [42] B. Simon and A. Zlatoš, *Higher-order Szego theorems with two singular points*, J. Approx. Theory *134* (2005), 114–129
- [43] S. Verblunsky, *On positive harmonic functions, I, II*, Proc. London Math. Soc. **38** (1935), 125–157, **40** (1936), 290–320.
- [44] P. Yuditskii, *Killip-Simon problem and Jacobi flow on GMP matrices*, Preprint. arXiv:1505.00972